

厚生労働科学研究費補助金 政策科学総合研究事業

(臨床研究等 ICT 基盤構築・人工知能実装研究事業)

分担研究報告書

リアルワールドデータの二次利用加速にむけた多施設データ処理方式の導入の試行研究

(大阪大学における検証)

研究分担者：岡田佳築 国立大学法人大阪大学 医学部附属病院 准教授

小西正三 国立大学法人大阪大学 医学部附属病院 特任助教（常勤）

研究要旨

わが国では AMED 事業において、臨床研究中核病院による RWD 利活用に向けた基盤構築の取組みである「臨中ネット」を進めており、各施設で共通の項目で構成された臨床研究のためのデータベース(臨中ネット共通 DB)を構築している。現状では、臨床研究を実施する研究者が自ら SQL を作成し、それを受け取ったデータ提供側の医療機関の担当者が自機関の臨中ネット共通 DB で SQL を実行することでデータ抽出を実施するが、データ抽出から受け渡しまで、研究者およびデータ抽出担当者にかかなりの負荷が生じることになる。本研究では、研究者がデータ提供元の各医療機関のデータベースに接続してデータ抽出が可能なデータ抽出システムを Google Cloud 上に構築し、効率的なデータ抽出および実際に本システムを用いる際の課題について検証することを目的としている。各施設からのデータ抽出については、システム内のソフトウェア(DTApp)を用いて抽出患者条件を設定して抽出する方法と、詳細条件を SQL で直接記載して抽出する方法を行ったが、前者ではデータ量が多い場合に抽出エラーになる場合があった。データ解析については、ARNI 内服開始前後の BNP・NT-proBNP の検査実施状況および検査結果の推移に関する解析を、解析に必要なデータの抽出をデータ抽出システムで行い解析は研究者の PC 上で行う方法と、解析に必要なデータの抽出、データの解析までを Google Cloud 上のデータ抽出システム上で行う方法で行った。共に、データ抽出にかかる労力と時間を大幅に削減することができると考えられたが、データの解析までを Google Cloud 上でおこなう方法では、現在、医療機関ごとの解析のみが可能であり、今後、複数の医療機関からのデータを統合した解析の実施が可能とする点が、特に課題になると考えられた。

A. 研究目的

わが国では AMED 事業において、臨床研究中核病院による RWD 利活用に向けた基盤構築の取組みである「臨中ネット」を進めている。本事業においては、各施設で共通の項目で構成された臨床研究のためのデータベース（臨中ネット

共通 DB）を構築している。臨中ネット共通 DB では、共通の項目でデータベースが構築されているため、データ抽出のための同じ SQL を用いて、各施設からデータ抽出が可能である。現状の計画では、臨床研究を実施する研究者が自ら SQL を作成し、それを受け取ったデータ提供側

の医療機関の担当者が自機関の臨中ネット共通DBでSQLを実行することでデータ抽出を実施するが、データ抽出から受け渡しまで、研究者および各施設のデータ抽出担当者にかかなりの負荷が生じることになる。上記の課題に対して、本研究ではGoogle Cloudが提供するモジュールを利用し、研究者がデータ提供元の各医療機関のデータベースに接続してデータ抽出が可能なデータ抽出システムを構築し、効率的なデータ抽出について検証するとともに、本システムを用いた解析を、システムの利用範囲を変更しながら試行的に実施し、実際に本システムを用いる際の課題について検証することを目的としている。

B. 研究方法

1. データ抽出システムを用いた各施設からのデータ抽出

まず、京都大学医学部附属病院を中心に構築したGoogle Cloud上のデータ抽出システムを、共同研究機関である、京都大学医学部附属病院、国立がん研究センター東病院、大阪大学医学部附属病院の臨中ネット共通DBに、各施設で接続するための環境構築を行った。本システムでは、Google Cloud上のデータ抽出・分析環境から、各施設の臨中ネット共通DBとODBD接続して、データ抽出が可能であるが、大阪大学医学部附属

病院では、臨中ネット共通DBが病院情報システムネットワーク内に設置されており、Google Cloudからの直接の接続・通信は、大阪大学医学部附属病院における情報セキュリティポリシーに抵触するため、大阪大学医学部附属病院の臨中ネット共通DBを複製したDBを、病院情報システムネットワーク外のセキュアな環境に構築し、データ抽出システムとの接続環境を構築した。

データ抽出に関しては、循環器疾患患者でアンジオテンシン受容体ネプリライシン阻害薬（ARNI）（商品名：エンレスト）が処方されている患者の、臨中ネット共通DBの対象症例IDリスト、検体検査結果テーブル、処方テーブルを抽出した。データ抽出システムを用いたデータ抽出については、システム内のソフトウェア（DTApp）を用いて、以下の2つの方法で行った。

- DTApp上で抽出患者条件を設定し、対象症例IDリスト、検体検査結果テーブル、処方テーブルを抽出（図1）
- SQLを直接記載し、対象症例IDリスト、検体検査結果テーブル、処方テーブルを個別に抽出（図2）

図1：DTApp画面（抽出患者条件を設定し抽出）

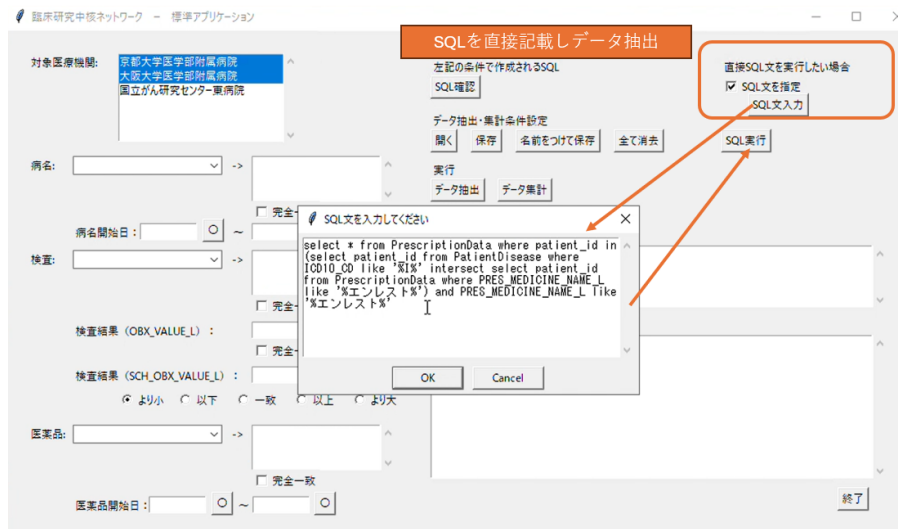


図 2 : DTApp 画面 (SQL を直接記載)

2. データ抽出システムの利用範囲を変更した解析の実施

BNP や NT-proBNP は心不全のバイオマーカーとしてその測定が推奨されている。ARNI はネブリライシンの作用を阻害することで、BNP の分解を抑制するため、ARNI 導入時には内因性の BNP が一時的に上昇する。そのため、ARNI 導入時にはバイオマーカーとしての BNP 値の解釈に注意が必要である。一方、NT-proBNP には ARNI による影響はなく、今回の試行的な解析として、ARNI 内服開始前後の BNP および NT-proBNP の検査実施状況および検査結果の推移に関する解析を行った。具体的な解析内容は以下とした。

対象患者：

- ARNI が 90 日以上処方された患者

解析項目：

- ARNI 処方日前後 90 日以内の BNP/NT-proBNP の測定状況
- ARNI 処方日前後 90 日以内の BNP/NT-proBNP 測定値の推移

Google Cloud 上のデータ抽出システムには、
「1. データ抽出システムを用いた各施設から

のデータ抽出」で用いたソフトウェアである DTApp に加えて、データ収集・データ分析・結果配信をすべてクラウド上で行うことができる DataTrain が実装されている (図3)。DataTrain はデータ収集 Train、データ分析 Train、結果配信 Train で構成され、それぞれ、以下の役割を担っている。

- データ収集 Train

- データ収集 Train に設定した SQL 文を用いて、各施設の臨中ネット共通 DB からデータテーブルを取り出し、CSV ファイルとしてクラウドに保存。

- データ分析 Train

- データ収集 Train が保存した CSV ファイルを取り出し、分析プログラムを実施し、分析結果をファイルとしてクラウドに保存。
- 分析プログラムは、Python もしくは R のプログラムが利用可能。

- 結果配信 Train

- データ分析 Train が保存した分析結果ファイルを、研究者がアクセス可能なクラウド領域に転送。

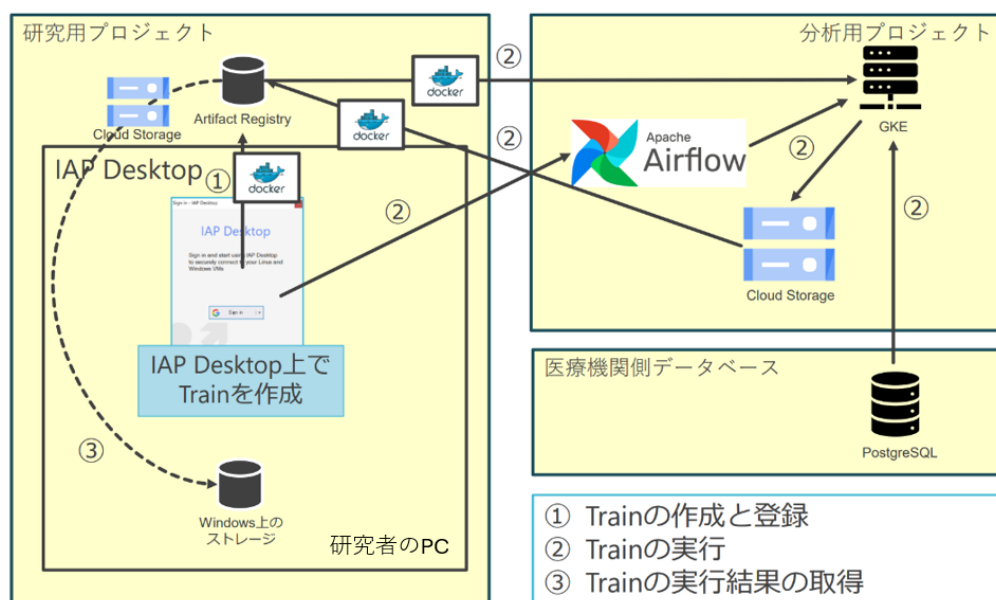


図 3 : DataTrain のイメージ図

上記のデータ抽出システムの機能を踏まえ、抽出システムの利用範囲としては、以下の 2 つのパターンを設定した。

- 解析に必要なデータの抽出をデータ抽出システムで行い、解析は研究者の PC 上で行う。

【具体的な解析手順】

- DTApp を用いて、解析に必要なデータテーブルを抽出
- 抽出したデータテーブルから、研究者の PC 上で解析用データシート作成
- 解析用データシートを、統計解析ソフトを用いて解析（統計解析ソフトは、グラフィカルユーザーインターフェース（GUI）操作での解析が可能な SPSS とコマンド入力が必要なキャラクターインターフェース(CUI)である R を用いて同じ解析を行う。）

- 解析に必要なデータの抽出、データの解析までを Google Cloud 上のデータ抽出システム上で行う。

【具体的な解析手順】

- 解析に必要なデータテーブルの抽出用 SQL をデータ収集 Train に設定
- 分析プログラムを R のプログラムとして作成し、データ分析 Train に設定
- DataTrain を実行
- 結果配信 Train により転送された分析結果ファイルをクラウド領域から研究者の PC へダウンロード

C. 研究結果

1. データ抽出システムを用いた各施設からのデータ抽出

【DTApp 上で抽出患者条件を設定し、対象症例 ID リスト、検体検査結果テーブル、処方テーブルを抽出】

DTApp 上で、病名の条件を循環器疾患（ICD10 コードが「I」ではじまる）および医薬品の条件を ARNI（薬品名が「エンレスト」）として、データ抽出を行った。DTApp の仕様では、この操作にて、臨中ネット共通 DB 上の、循環器疾患の病名登録があり、かつ、ARNI の処方がある患者の対象症例 ID リストと、検体検査結果テーブル、処方テーブル内のすべてのデータが抽出される。

大学医学部附属病院、京都大学医学部附属病院、国立がん研究センター東病院から抽出された対象症例 ID リストの ID 件数、検体検査結果テーブルのデータ件数、処方テーブルのデータ件数を表 1 に示す。

表 1：抽出されたデータ件数（DTApp 上で抽出患者条件設定）

	阪大病院	京大病院	国がん東病院
対象症例 ID リスト ID 件数	433	853	65
検体検査結果テーブルデータ件数	975,404	3,231,326	69,353
処方テーブルデータ件数	218,564	340,671	5,339

検体検査結果テーブルについては、施設によっては対象のデータ数が非常に大きくなり、抽出時のリソースが不足し、抽出エラーとなるケースが見受けられた。そのような場合は、医薬品開始日を限定し、数回に分けてデータを抽出することで、データの抽出が可能であった。

【SQL を直接記載し、対象症例 ID リスト、検体検査結果テーブル、処方テーブルを抽出】

SQL 文を直接記載する方法では、より詳細な抽出条件で、対象症例 ID リスト、検体検査結果テーブル、処方テーブルをそれぞれ抽出可能であるため、データ件数の多かった、大阪大学医学部附属病院および京都大学医学部附属病院の DB から、次の「2. データ抽出システムの利用範囲を変更した解析の実施」の解析テーマで必要な範囲のデータに絞り抽出を行った。具体的には、検体検査結果テーブルは BNP および NT-proBNP のデータのみ、処方テーブルは ARNI のデータのみ、以下の SQL 文を用いて抽出を行った。

- 対象症例 ID リストの SQL
 - `select patient_id from PatientDisease where ICD10_CD like '%I%' intersect select patient_id from PrescriptionData where PRES_MEDICINE_NAME_L like '%エンレスト%'`
- 検体検査結果テーブルの SQL
 - `select * from ObservationResult where patient_id in (select patient_id from PatientDisease where ICD10_CD like '%I%' intersect select patient_id from PrescriptionData where PRES_MEDICINE_NAME_L like '%エンレスト%') and OBX_NAME_L in (select OBX_NAME_L from ObservationResult where OBX_NAME_L like '%BNP%' or OBX_NAME_L like '%NT-proBNP%')`
- 処方テーブルの SQL
 - `select * from PrescriptionData where patient_id in (select patient_id from PatientDisease where ICD10_CD like '%I%' intersect select patient_id from PrescriptionData where PRES_MEDICINE_NAME_L like '%エンレスト%')`

抽出されたデータ件数は、表 2 に示す通り大幅に減少し、抽出時のリソースが不足等による抽出エラーを認めることなく抽出が可能であった。

表 2：抽出されたデータ件数（SQL を直接記載）

	阪大病院	京大病院
対象症例 ID リスト ID 件数	433	853
検体検査結果テーブルデータ件数	7,082	15,320
処方テーブルデータ件数	2,990	8,953

2. データ抽出システムの利用範囲を変更した解析の実施

【解析に必要なデータの抽出をデータ抽出システムで行い、解析は研究者の PC 上で行う。】

解析に必要なデータの抽出は、件数の多い大阪大学医学部附属病院および京都大学医学部附属病院から、「1. データ抽出システムを用いた各施設からのデータ抽出」の SQL 文を直接記載する方法で行った条件で、ARNI が処方されている症例の検体検査結果テーブル（BNP および NT-proBNP のデータのみ）と、処方テーブル（ARNI のみ）を抽出した。

上記で抽出したテーブルから、対象患者を ARNI が 90 日以上処方された患者を対象に、解析項目である、ARNI 処方日前後 90 日以内の BNP/NT-proBNP の測定状況と ARNI 処方日前後 90 日以内の BNP/NT-proBNP 測定値（複数回測定の場合は平均値）の推移を解析するためのデータテーブルを R のプログラムを用いて作成し、解析用のデータテーブルを用いて、統計ソフトの SPSS および R のプログラムを用いて解析を実施した。

表 3 に ARNI が 90 日以上処方された患者の、ARNI 処方日前後 90 日以内の BNP/NT-proBNP の測定状況を示す。（数値は、SPSS による解析、R による解析ともに同じ数値である。）

表 3：BNP/NT-proBNP の測定状況

	阪大病院	京大病院
N	161	129
pattern (%)		
fromBNP_toBNP	30 (18.6)	31 (24.0)
fromBNP_toNT-proBNP	11 (6.8)	9 (7.0)
fromNT-proBNP_toBNP	0 (0)	6 (4.7)
fromNT-proBNP_toNT-proBNP	61 (37.9)	17 (13.2)
no_post	10 (6.2)	15 (11.6)
no_pre	49 (30.4)	51 (39.5)

ARNI が 90 日以上処方された患者は、大阪大学医学部附属病院で 161 症例（症例の処方開始日の範囲は 2020 年 9 月 25 日から 2022 年 12 月 26 日）で、京都大学医学部附属病院で 129 症例（症例の処方開始日の範囲は 2021 年 1 月 19 日から 2025 年 2 月 25 日）であった。処方日前 90 日以内および処方日後 90 日以内の両方に、BNP もしくは NT-proBNP の測定があった症例は、大阪大学医学部附属病院では 102 症例であり、京都大学医学部附属病院では 63 症例であった。このうち ARNI 開始前に BNP を測定していた症例は大阪大学医学部附属病院で 41 症例であり、ARNI 開始後に NT-proBNP に変更した症例（fromBNP_toNT-proBNP）は 11 症例、引き続き BNP を測定した症例（fromBNP_toBNP）は 30 症例であった。京都大学医学部附属病院で ARNI 開始前に BNP を測定していた症例は、40 症例で、NT-

proBNP に変更した症例は 9 症例、引き続き BNP を測定した症例は 31 症例であった。ARNI 開始前に BNP を測定していた多くの症例は、そのまま BNP が測定されていた。

ARNI 処方日前後 90 日以内の BNP/NT-proBNP 測定値の推移については、処方日前後に BNP の測定がある症例および NT-proBNP の測定がある症例を対象として解析を行った。解析については、平均値と標準偏差を算出し、箱ひげ図でグラフ表示を行い、前後 90 日以内の測定値の推移については、対応のある t 検定にて検定（有意水準：0.05）を行った。結果を図 4 に示しているが、大阪大学医学部附属病院、京都大学医学部附属病院ともに ARNI 開始後、NT-proBNP は有意に低下を示していたが、BNP は有意な低下は認めていなかった。

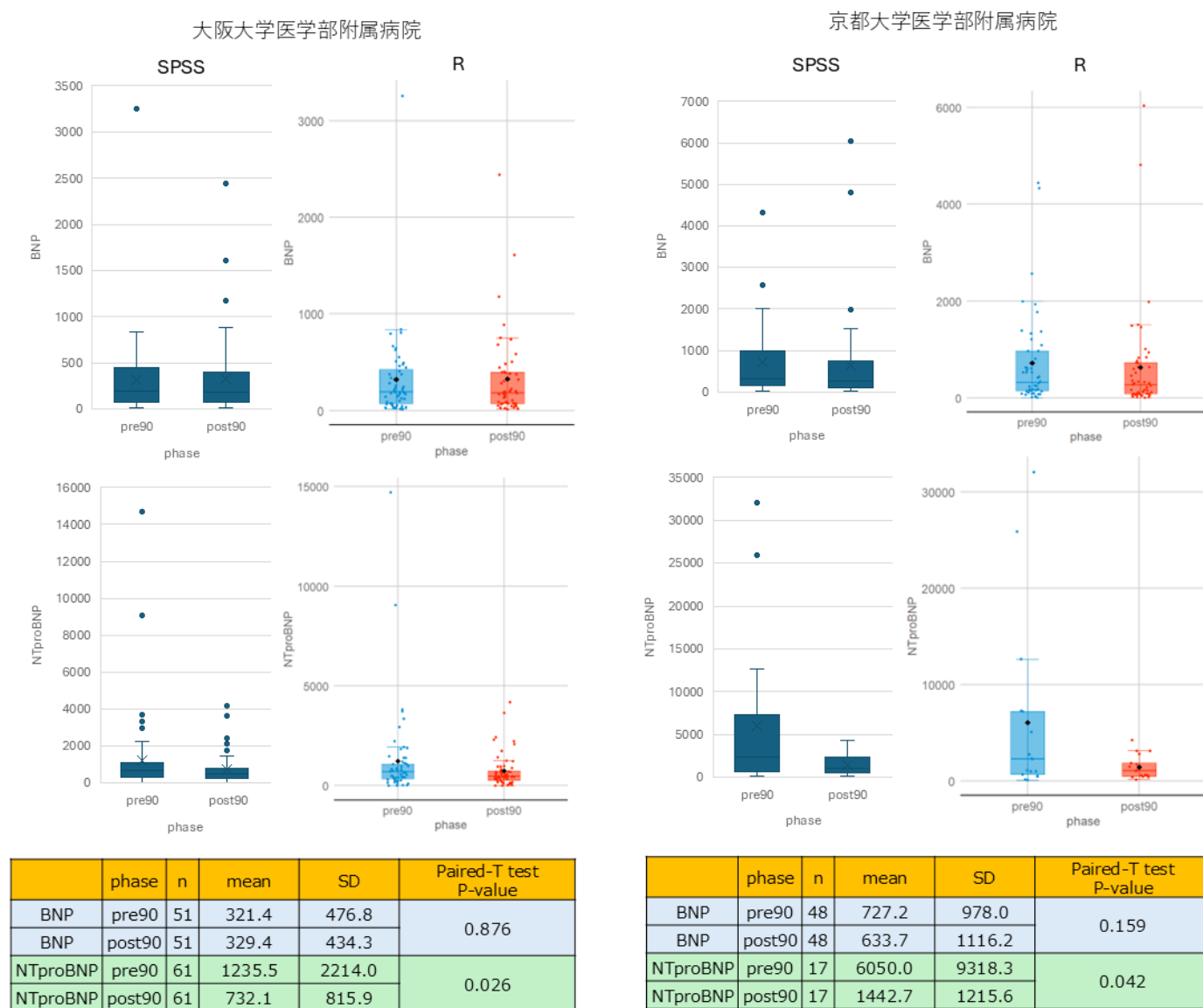


図 4：ARNI 投与前後の BNP/NT-proBNP 値

【解析に必要なデータの抽出、データの解析までを Google Cloud 上のデータ抽出システム上で行う。】

データ収集 Train に設定する、解析に必要なデータテーブルの抽出用 SQL は、「1. データ抽出システムを用いた各施設からのデータ抽出をデータ収集 Train に設定」の「SQL を直接記載し、対象症例 ID リスト、検体検査結果テーブル、処方テーブルを抽出」の部分で用いた、SQL を用いた。

データ分析 Train に設定する分析プログラムは、本セクションの前半の「解析に必要なデータの抽出をデータ抽出システムで行い、解析は研究者の PC 上で行う。」部分で作成・使用した、解析用データテーブルの作成および統計解析用の R のプログラムを流用した。プログラムの流用については、R のプログラム上のファイル読み込みおよび書き出しのファイルパスを指定する部分を、Google Cloud のデータ抽出システムの構成に合わせたパスに修正することで流用可能であった。

上記にて設定を行った DataTrain を実行した後に、結果配信 Train により転送された分析結果ファイルをクラウド領域から研究者の PC へダウンロードして、分析結果を確認した所、本セクションの前半の「解析に必要なデータの抽出をデータ抽出システムで行い、解析は研究者の PC 上で行う。」部分で R を用いて行った、表 3・図 1 と同じ、解析結果と箱ひげ図が、分析結果としてダウンロードされていた。

D. 考察

1. データ抽出システムを用いた各施設からのデータ抽出

データ抽出については、DTApp 上で条件を指定することでデータ抽出が可能であったが、検体検査結果テーブル、処方テーブルについては、対象症例のすべてのデータが抽出されるため、葬礼数によっては、非常に膨大なデータ量になり、データ抽出がエラーとなり抽出できないケースが認められた。そのため、必要な項目を絞った形で抽出が可能な、SQL を直接記載しての抽出のほうが、解析に向けたデータ抽出においては、データ抽出がスムーズにいくものと考えられた。

SQL を直接記載しての抽出では、SQL を研究者が作成する必要があるが、DTApp では、DTApp 上で条件を指定した際に DTApp 上で作成される SQL を

確認することが可能であり、この SQL を土台として、必要な SQL を作成することで、比較的 SQL 作成は行いやすいものと考えられた。

2. データ抽出システムの利用範囲を変更した解析の実施

最初に行った、解析に必要なデータの抽出をデータ抽出システムで行い、解析は研究者の PC 上で行う方法については、研究者が一般的に行う解析手順になると考えられる。また、今回の解析項目を問題なく評価することが可能であった。

従来の臨中ネット共通 DB における解析手順と比べると、同一の SQL を用いて、クラウド上で必要なデータ抽出を行うことができるため、データ抽出にかかる労力および時間は、大幅に軽減されているものと考えられる。また、データ抽出後は、グラフィカルユーザーインターフェース (GUI) 操作での解析が可能な SPSS とコマンド入力が必要なキャラクターユーザーインターフェース (CUI) である R を用いて同様の結果が出ており、普段研究者が利用している統計解析ソフトによる解析につなげることが可能であると考えられる。

ただ、データ抽出については、検体検査結果テーブルや処方テーブルなどのデータテーブルは、別々に抽出されるため、これらのテーブルから解析用のデータテーブルを作成する必要がある。この解析用データテーブルの作成については、今回 R のプログラムを用いて、スムーズに作成することが可能であったが、R や Python のプログラムを普段利用していない研究者が解析用データテーブルを作成する場合には、この部分には労力がかかる可能性があると考えられた。

後半に行った、解析に必要なデータの抽出、データの解析までを Google Cloud 上のデータ抽出システム上で行う方法については、普段の解析で R や Python でプログラムを作成して解析を行っている場合には、比較的容易に行うことが可能であった。そのため、本システムを用いて、解析までをすべてクラウド上で実施することが可能であり、データ解析の労力は、大幅に軽減できるものと考えられた。しかしながら、現在のシステムの仕様では、医療機関ごとの解析しか実施できず、複数の医療機関からのデータを統合したデータを用いた解析の実施については、データ抽出後に研究者が自身の PC 上で統合したデータを作成して解析を行う必要がある。今後、

本システムを用いた臨床研究を推進していくためには、複数の医療機関からのデータを統合した解析の実施が行うことができるよう、システムを更新していくことが必要であると考えられた。

E. 結論

Google Cloud 上のデータ抽出システムを用いることで、複数医療機関の臨中ネット共通 DB から、効率的にデータ抽出およびデータ解析が可能であった。より効率的な実施には、システムに実装されている DTApp や DataTrain などの機能を、データ収集・解析の目的に応じて、活用することが重要であり、今後、本システムを用いた臨床研究を推進していくためには、複数の医療機関からのデータを統合した解析の実施が行うことができるよう、システムを更新

していくことが、特に課題になると考えられた。

F. 健康危険情報

なし

G. 研究発表

1. 論文発表

なし

2. 学会発表

なし

H. 知的財産権の出願・登録状況

1. 特許取得

なし

2. 実用新案登録

なし

3. その他

なし